Exploratory Data Analysis

Lecture 1

Corina Besliu

Technical University of Moldova

September 6, 2021

Outline

- Subject of Research & Methods
- Motivation
- Concepts and Definitions
- Practical Assignments in Python

TECHNICAL UNIVERSITY OF MOLDOVA

Outline

- Subject of Research & Methods
- Motivation
- Concepts and Definitions
- Practical Assignments in Python

TECHNICAL UNIVERSITY OF MOLDOVA

Outline

- Subject of Research & Methods
- Motivation
- Concepts and Definitions
- Practical Assignments in Python

TECHNICAL UNIVERSITY OF MOLDOVA

Outline

- Subject of Research & Methods
- Motivation
- Concepts and Definitions
- Practical Assignments in Python

TECHNICAL UNIVERSITY OF MOLDOVA

Lecture Outline

TECHNICAL UNIVERSITY OF MOLDOVA

Subject of Research

Subject of Research

Exploratory data analysis deals with methods of organizing, summarizing, and presenting data in a convenient and informative way to allow statistics practitioners present data in ways that make it easy for the reader to extract useful information.

Methods

Methods

- Descriptive Statistics

  Descriptive Statistics are coefficients that summarize a given data set.

  - central tendency: mean, median, mode
  - variability: variance, standard deviation, range, kurtosis, skewness

- Graphical Techniques

  Graphical summaries are great tool to communicate information, because people prefer to look at pictures rather than at numbers.

  pie charts, bar graphs, histograms, box plots, dot plots, scatter plots

Methods

Methods

- Descriptive Statistics

  Descriptive Statistics are coefficients that summarize a given data set.

  - central tendency: mean, median, mode
  - variability: variance, standard deviation, range, kurtosis, skewness

- Graphical Techniques

  Graphical summaries are great tool to communicate information, because people prefer to look at pictures rather than at numbers.

  pie charts, bar graphs, histograms, box plots, dot plots, scatter plots

Methods

Methods

- Descriptive Statistics

  Descriptive Statistics are coefficients that summarize a given data set.

  - central tendency: mean, median, mode
  - variability: variance, standard deviation, range, kurtosis, skewness

- Graphical Techniques

  Graphical summaries are great tool to communicate information, because people prefer to look at pictures rather than at numbers.

  pie charts, bar graphs, histograms, box plots, dot plots, scatter plots

Methods

Methods

- Descriptive Statistics

  Descriptive Statistics are coefficients that summarize a given data set.

  - central tendency: mean, median, mode
  - variability: variance, standard deviation, range, kurtosis, skewness

- Graphical Techniques

  Graphical summaries are great tool to communicate information, because people prefer to look at pictures rather than at numbers.

  pie charts, bar graphs, histograms, box plots, dot plots, scatter plots

Methods

Methods

- Descriptive Statistics

  Descriptive Statistics are coefficients that summarize a given data set.

  - central tendency: mean, median, mode
  - variability: variance, standard deviation, range, kurtosis, skewness

- Graphical Techniques

  Graphical summaries are great tool to communicate information, because people prefer to look at pictures rather than at numbers.

  pie charts, bar graphs, histograms, box plots, dot plots, scatter plots

Lecture Outline

Motivation (slide 1)

Why are descriptive statistics important?

In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29 F at launch.

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F): 66, 70, 69, 80, 68, 67, 72, 70, 70, 57, 63, 70, 78, 67, 53, 67, 75, 70, 81, 76, 79, 75, 76, 58, 29.

Motivation (slide 1)

Why are descriptive statistics important?

In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29 F at launch.

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F): 66, 70, 69, 80, 68, 67, 72, 70, 70, 57, 63, 70, 78, 67, 53, 67, 75, 70, 81, 76, 79, 75, 76, 58, 29.

Motivation (slide 1)

Why are descriptive statistics important?

In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29 F at launch.

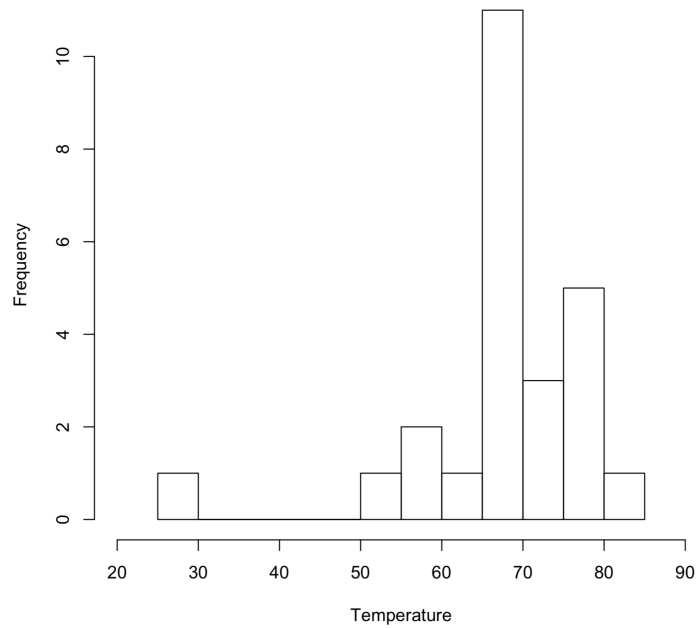
Here are the launch-temperatures of the first 25 shuttle missions (in degrees F): 66, 70, 69, 80, 68, 67, 72, 70, 70, 57, 63, 70, 78, 67, 53, 67, 75, 70, 81, 76, 79, 75, 76, 58, 29.

Motivation (slide 1)

Why are descriptive statistics important?

In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29 F at launch.

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F): 66, 70, 69, 80, 68, 67, 72, 70, 70, 57, 63, 70, 78, 67, 53, 67, 75, 70, 81, 76, 79, 75, 76, 58, 29.



Motivation (slide 2)

The two most important functions of descriptive statistics are:

- Communicate information.
- Support reasoning about data.

When exploring data of large size, it becomes essential to use summaries.

Motivation (slide 2)

The two most important functions of descriptive statistics are:

- Communicate information.
- Support reasoning about data.

When exploring data of large size, it becomes essential to use summaries.

Motivation (slide 2)

The two most important functions of descriptive statistics are:

- Communicate information.
- Support reasoning about data.

When exploring data of large size, it becomes essential to use summaries.

Motivation (slide 2)

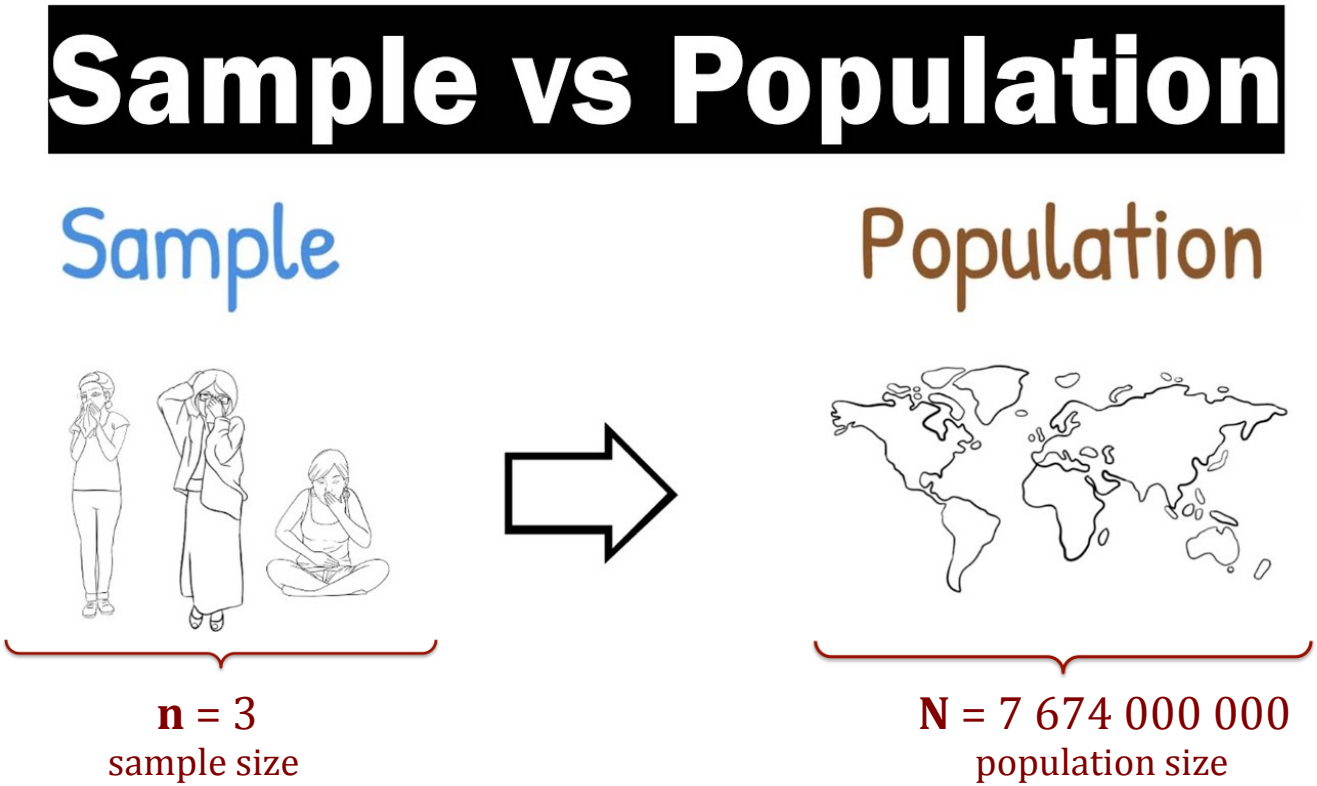
The two most important functions of descriptive statistics are:

- Communicate information.
- Support reasoning about data.

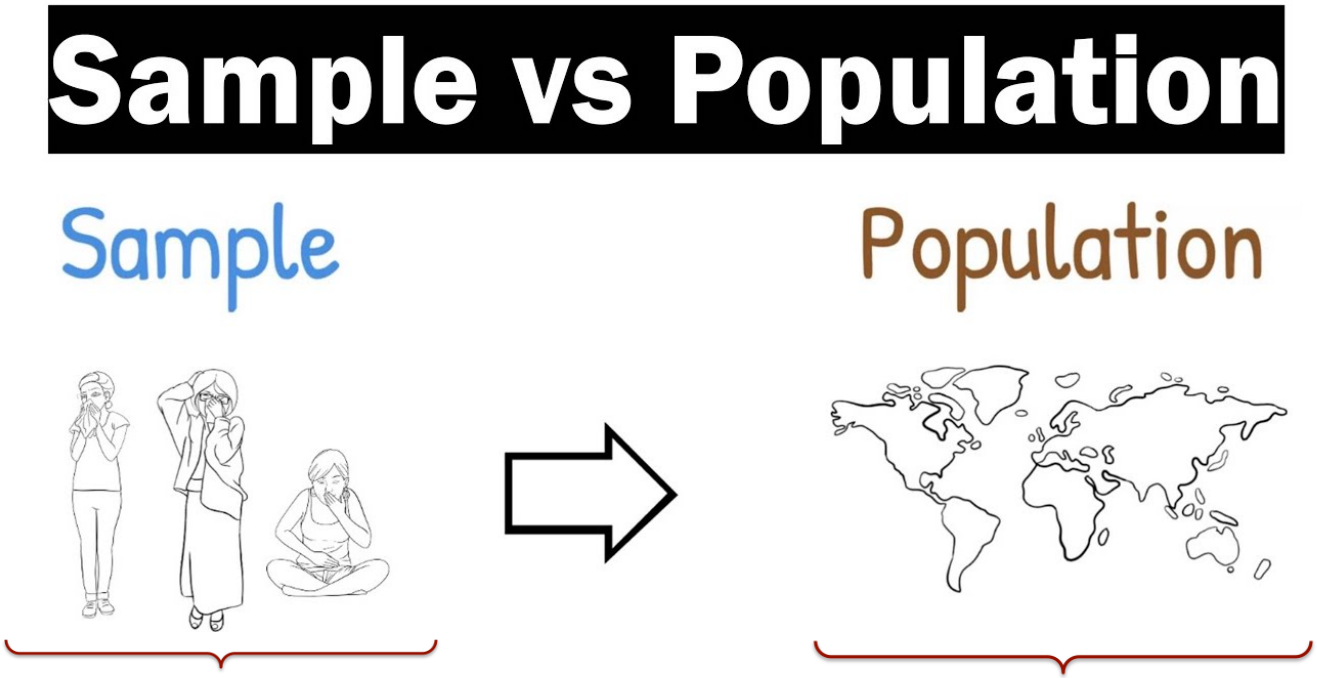
When exploring data of large size, it becomes essential to use summaries.

Lecture Outline

Sample and Population



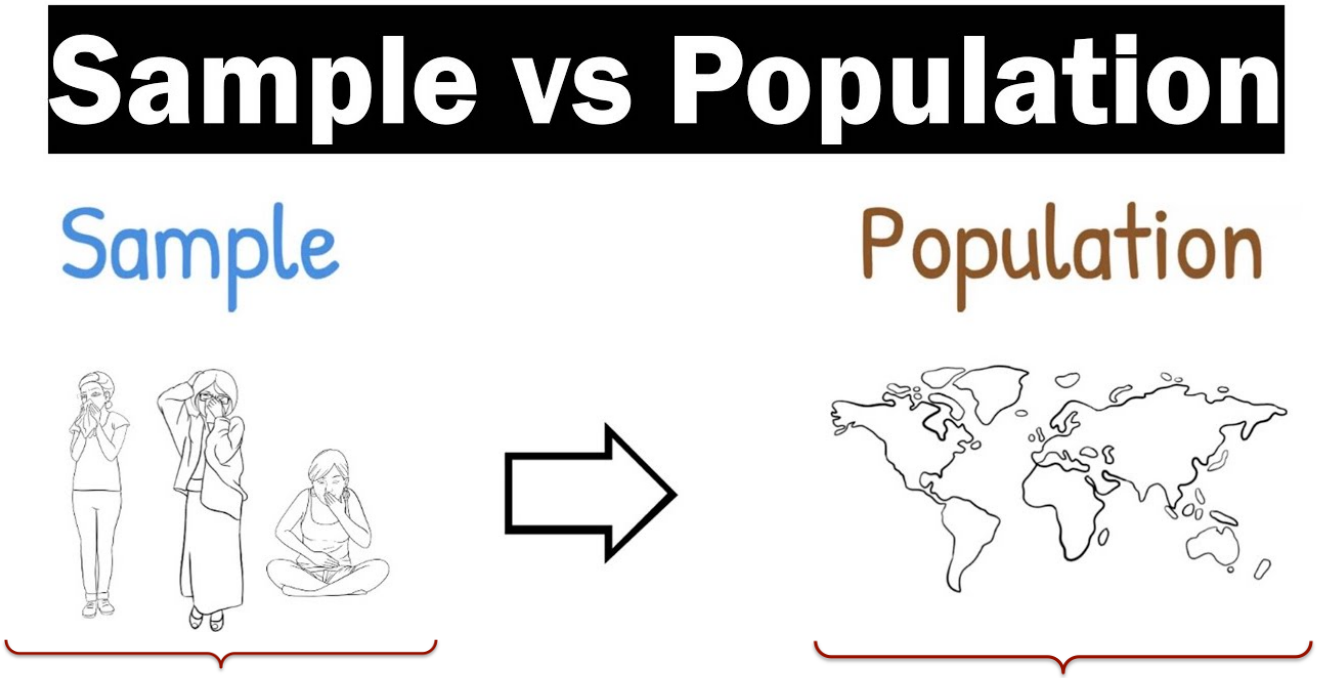
Sample and Population



Population is a Relative Concept

People in R. Moldova **N** = 4 023 039

Sample and Population



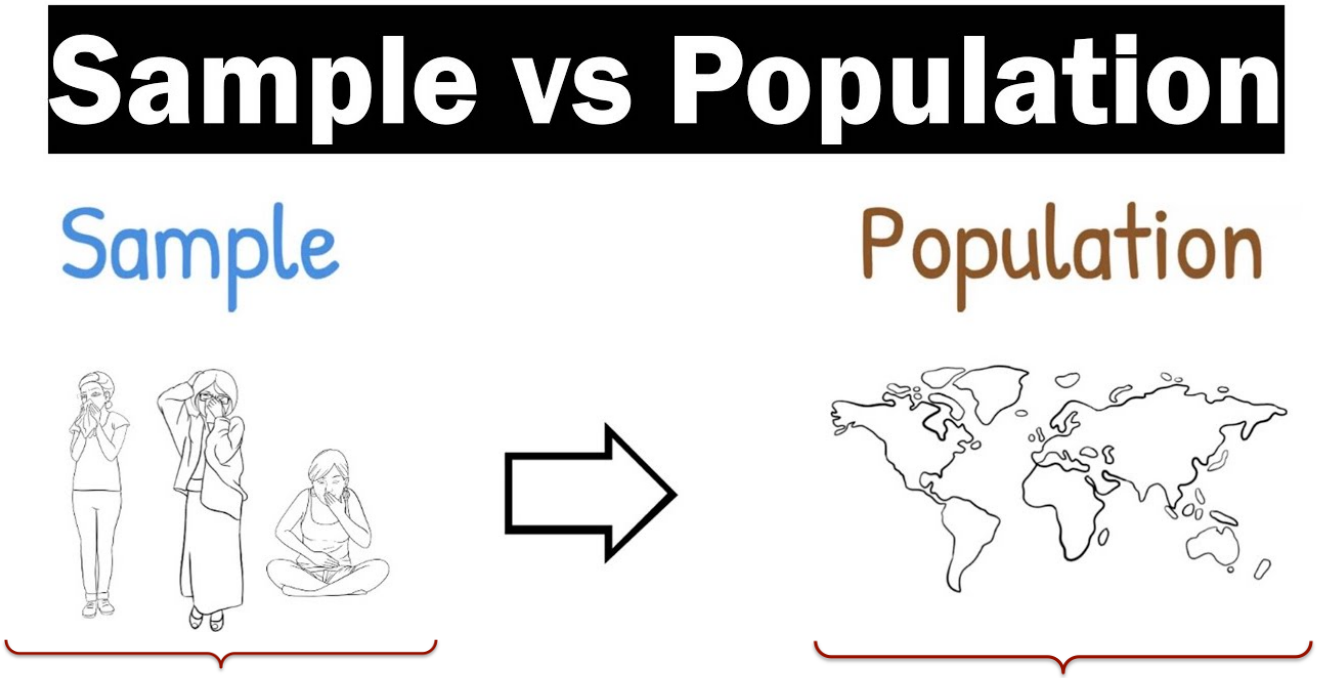
n = 3
sample size

N = 7 674 000 000
population size

Population is a Relative Concept

People in R. Moldova	**N** = 4 023 039
Pigs in the world	**N** = 677 600 000

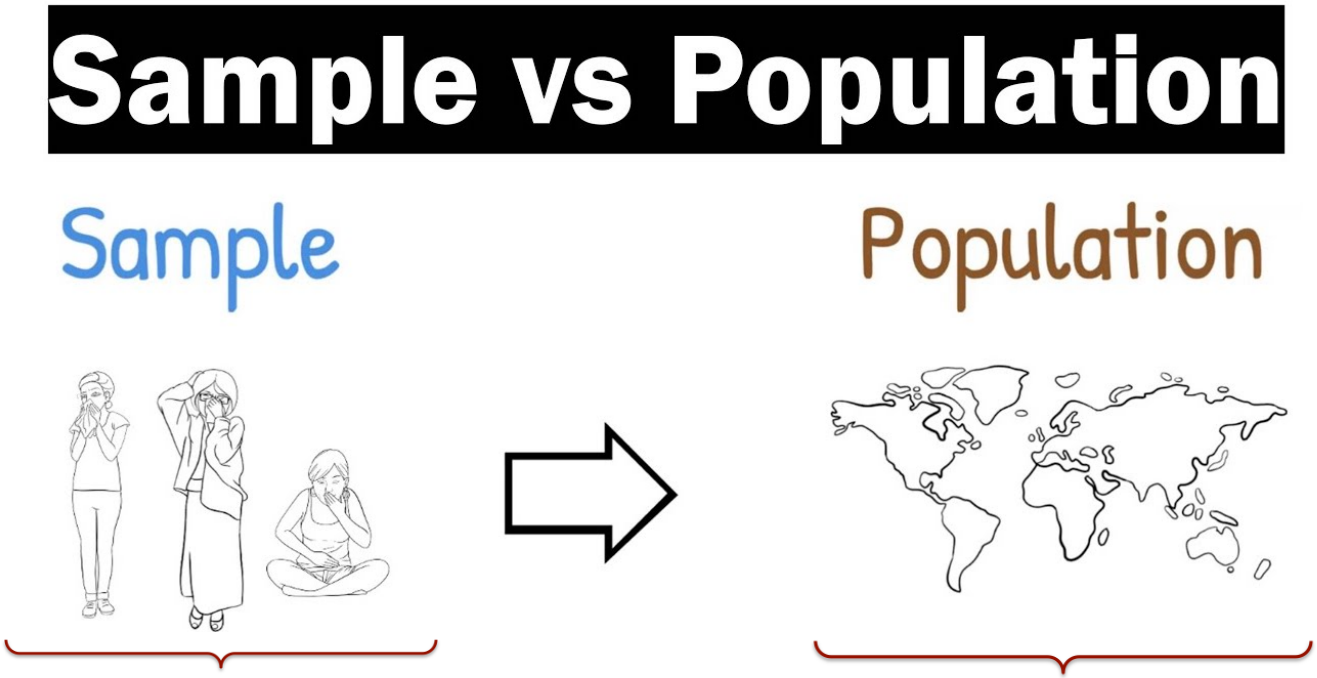
Sample and Population



Population is a Relative Concept

People in R. Moldova	**N** = 4 023 039
Pigs in the world	**N** = 677 600 000
Computers in the world	**N** = 2 000 000 000

Sample and Population



Population is a Relative Concept

n = ?	People in R. Moldova	**N** = 4 023 039
n = ?	Pigs in the world	**N** = 677 600 000
n = ?	Computers in the world	**N** = 2 000 000 000

Definitions

Population

In statistics, a population is the entire group that you want to draw conclusions about and from which statistical samples are drawn to be studied. It need not consist only of people. We can have population of heights, weights, BMIs, hemoglobin levels, events, outcomes, etc.

Sample

A sample is a subset containing the characteristics of a larger population that you will collect data from. Samples are used in statistical testing when population sizes are too large for the test to include all possible members or *observations*.

$$\mathbf{n < N}$$

Definitions

Population

In statistics, a population is the entire group that you want to draw conclusions about and from which statistical samples are drawn to be studied. It need not consist only of people. We can have population of heights, weights, BMIs, hemoglobin levels, events, outcomes, etc.

Sample

A sample is a subset containing the characteristics of a larger population that you will collect data from. Samples are used in statistical testing when population sizes are too large for the test to include all possible members or *observations*.

$$n < N$$

Definitions

Population

In statistics, a population is the entire group that you want to draw conclusions about and from which statistical samples are drawn to be studied. It need not consist only of people. We can have population of heights, weights, BMIs, hemoglobin levels, events, outcomes, etc.

Sample

A sample is a subset containing the characteristics of a larger population that you will collect data from. Samples are used in statistical testing when population sizes are too large for the test to include all possible members or *observations*.

$$\mathbf{n} < \mathbf{N}$$

Observations

Observations

In statistics, an observation is one occurrence of something you are measuring.

- The total number of observations is equal to the sample size of the dataset.
- A single observation can be associated with multiple variables.

Observations

Observations

In statistics, an observation is one occurrence of something you are measuring.

- The total number of observations is equal to the sample size of the dataset.
- A single observation can be associated with multiple variables.

Observations

Observations

In statistics, an observation is one occurrence of something you are measuring.

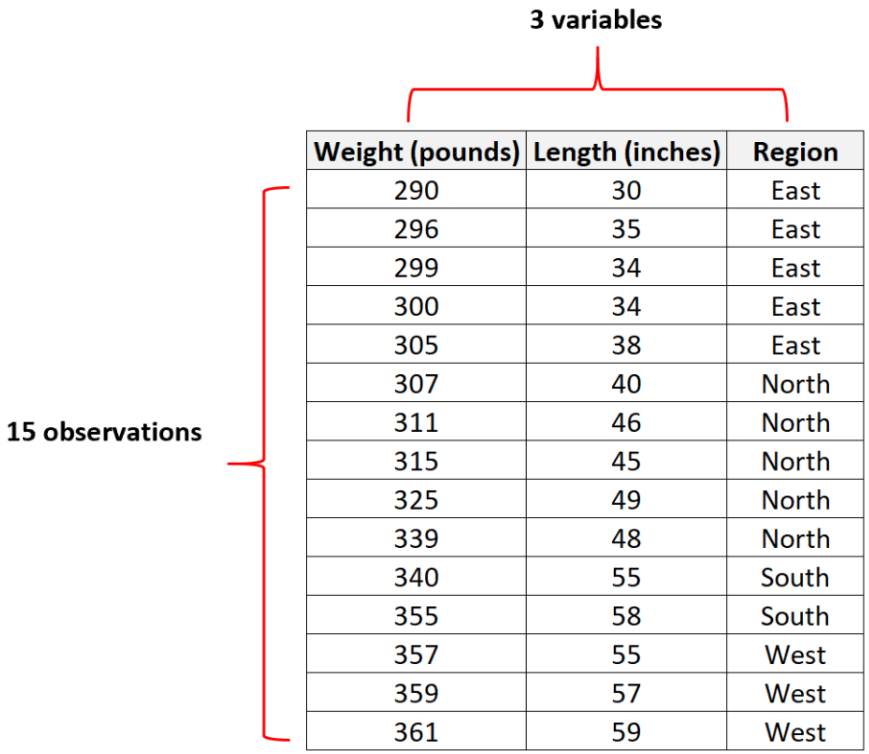
- The total number of observations is equal to the sample size of the dataset.
- A single observation can be associated with multiple variables.

Observations

Observations

In statistics, an observation is one occurrence of something you are measuring.

- The total number of observations is equal to the sample size of the dataset.
- A single observation can be associated with multiple variables.



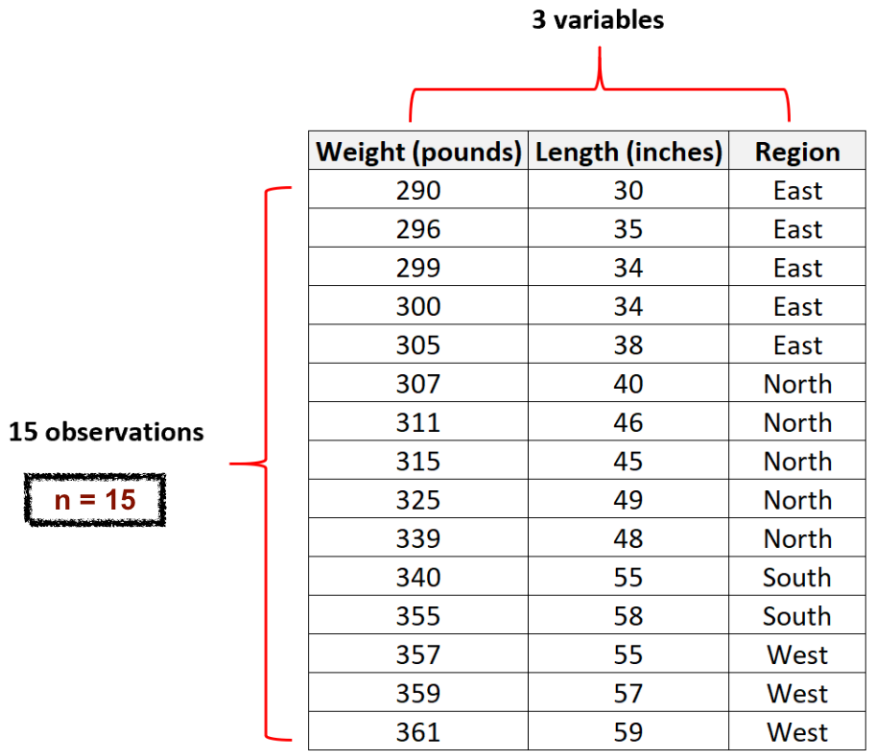
Weight (pounds)	Length (inches)	Region
290	30	East
296	35	East
299	34	East
300	34	East
305	38	East
307	40	North
311	46	North
315	45	North
325	49	North
339	48	North
340	55	South
355	58	South
357	55	West
359	57	West
361	59	West

Observations

Observations

In statistics, an observation is one occurrence of something you are measuring.

- The total number of observations is equal to the sample size of the dataset.
- A single observation can be associated with multiple variables.



Weight (pounds)	Length (inches)	Region
290	30	East
296	35	East
299	34	East
300	34	East
305	38	East
307	40	North
311	46	North
315	45	North
325	49	North
339	48	North
340	55	South
355	58	South
357	55	West
359	57	West
361	59	West

Lecture Outline

The Mean

Mean

In statistics, the mean or the average is a measure of central tendency of a probability distribution. It is also referred to as an expected value.

- Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mean Properties

Mean Properties

- Meaningful for interval and ratio data (continuous variables).
- Affected by unusually large or small observations (outliers).
  - Hence median is also useful

Mean Properties

Mean Properties

- Meaningful for interval and ratio data (continuous variables).
- Affected by unusually large or small observations (outliers).
  - Hence median is also useful

Mean Properties

Mean Properties

- Meaningful for interval and ratio data (continuous variables).
- Affected by unusually large or small observations (outliers).
  - Hence median is also useful

Mean Properties

Mean Properties

- Meaningful for interval and ratio data (continuous variables).
- Affected by unusually large or small observations (outliers).
  - Hence median is also useful
- The only measure of central tendency where the sum of the deviation of each value from the measure is zero, i.e.,

$$\sum(x_i - \bar{x}) = 0$$

Median Properties

Median

Median corresponds to the middle value when data are ordered from smallest to largest value (min to max). This results in an equal number of observations above the median as bellow it.

22.000 26.000 26.000 30.000 38.000 40.000 42.000 42.000 42.000 180.000 200.000

Properties

- Not affected by extreme values.
- Meaningful for ratio, interval, and ordinal data.

Median Properties

Median

Median corresponds to the middle value when data are ordered from smallest to largest value (min to max). This results in an equal number of observations above the median as bellow it.

22.000 26.000 26.000 30.000 38.000 40.000 42.000 42.000 42.000 180.000 200.000

Properties

- Not affected by extreme values.
- Meaningful for ratio, interval, and ordinal data.

Median Properties

Median

Median corresponds to the middle value when data are ordered from smallest to largest value (min to max). This results in an equal number of observations above the median as bellow it.

22.000 26.000 26.000 30.000 38.000 40.000 42.000 42.000 42.000 180.000 200.000

Properties

- Not affected by extreme values.
- Meaningful for ratio, interval, and ordinal data.

Median Properties

Median

Median corresponds to the middle value when data are ordered from smallest to largest value (min to max). This results in an equal number of observations above the median as bellow it.

22.000 26.000 26.000 30.000 38.000 40.000 42.000 42.000 42.000 180.000 200.000

Properties

- Not affected by extreme values.
- Meaningful for ratio, interval, and ordinal data.

Mode Properties

Mode

The mode is the observation that occurs most frequently

22.000 26.000 26.000 30.000 38.000 40.000 42.000 42.000 42.000 180.000 200.000

Properties

- Useful when data consist of a small number of unique values.
- Can be used when the data are nominal or categorical such as gender, religion, political affiliation, etc.

Mode Properties

Mode

The mode is the observation that occurs most frequently

22.000 26.000 26.000 30.000 38.000 40.000 42.000 42.000 42.000 180.000 200.000

Properties

- Useful when data consist of a small number of unique values.
- Can be used when the data are nominal or categorical such as gender, religion, political affiliation, etc.

Mode Properties

Mode

The mode is the observation that occurs most frequently

22.000 26.000 26.000 30.000 38.000 40.000 42.000 42.000 42.000 180.000 200.000

Properties

- Useful when data consist of a small number of unique values.
- Can be used when the data are nominal or categorical such as gender, religion, political affiliation, etc.

Mode Properties

Mode

The mode is the observation that occurs most frequently

22.000 26.000 26.000 30.000 38.000 40.000 42.000 42.000 42.000 180.000 200.000

Properties

- Useful when data consist of a small number of unique values.
- Can be used when the data are nominal or categorical such as gender, religion, political affiliation, etc.

Lecture Outline

Let's get Started!

Access Google Colaboratory

Create a Gmail account at: `https://accounts.google.com/signup`